SEJODR

StaTips Part V: The adjustment of the P value in the context of multiple comparisons

Perinetti, Giuseppe *

** Private practice, Nocciano (PE), Italy*

ABSTRACT

Very often researchers face the problem of multiple comparisons using the same data sets. When performing multiple comparisons, it happens that some P values below the desired cut-off level will be retrieved by chance, thus leading the researcher to have false positives. To reduce the risk of making wrong conclusions, statistical adjustment of the P values in the context of multiple comparisons has been proposed over the last decades. This article will deal with two simple procedures, Bonferroni and Holm, commonly used in dental literature to adjust the P values in case of multiple comparisons.

Perinetti G. StaTips Part V: The adjustment of the P value in the context of multiple comparisons. South Eur J Orthod Dentofac Res. 2018;5(2):3-4.

FRAMING OF THE PROBLEM

Very often researchers face the problem of multiple comparisons, using the same data sets. An example may be the case of a cross-sectional comparison of any variable, among at least three different groups of patients (called A, B and C). As stated in i StaTips Part I[1], a first step would be to test the significance of the difference in the interested variable among the groups (by means of an analysis of variance or other methods). If a significant difference is retrieved, the next step would be to perform the multiple comparisons between the groups, i.e. Group A vs. Group B, Group A vs. Group C and Group B vs. Group C. In the case of three groups, a total of 3 multiple comparisons will be needed; however, in the case of four groups, the total number of multiple comparisons raises to 6 and so on. The same applies to the case of longitudinal data sets where data is collected within the same group in more than 2 sessions.

However, when performing multiple comparisons, it happens that some P values below the customary cut-off level (0.05) will be retrieved by chance, thus leading the researcher to a wrong conclusion that a given treatment had an effect when it was not the case (false positives or Type I error). Specifically, the probability that one or more Type I errors will occur in case of multiple comparisons is referred to as 'family-wise error rate'. To reduce the family-wise error rate, and thus the risk of making wrong conclusions, statistical adjustment of the P values in the context of multiple comparisons has been proposed over the last decades[2-4]. In spite of the relevance of the issue, there is still no universally accepted approach for the adjustment of the P value, which may not be a trivial task. Indeed, a conservative approach may be protective against false positives (Type I error) but may produce false negatives (Type II errors) and *vice-versa*. This article will deal with two simple procedures, Bonferroni and Holm, commonly used in dental literature to adjust the P values in the context of multiple comparisons. For other procedures, see Shafer.[4]

Corresponding Author:
Perinetti Giuseppe
Via San Lorenzo 69/1,
65010 Nocciano (PE), Italy.
e-mail: G.Perinetti@yahoo.com

BONFERRONI PROCEDURE

The Bonferroni procedure is likely the simplest and most widely used method for the adjustment of the P values for multiple comparisons. In this single-step procedure, the all adjusted P values for a batch of multiple comparisons are chosen to be equal, and the method is then called the unweighted Bonferroni procedure. For the overall cut-off level of significance (usually 0.05) and n multiple comparisons the Bonferroni procedure is as follows:

$$\frac{\textit{Overall cut-off level of P}}{n}$$

Where:

- Overall cut-off level of P, usually 0.05
- n = number of multiple comparisons.

In the case of 3 multiple comparisons, then the adjusted P value would be 0.05/3=0.0167; however, in the case of 20 comparisons the adjusted P value would drop to 0.002 possibly leading to false negative results. In spite the Bonferroni is simple to calculate, it then suffers a lack of statistical power and should be avoided in case of more than 8-10 multiple comparisons[2]. A worked example with 6 multiple comparisons is also reported in Table 1 where, after the Bonferroni procedure, only 1 comparison is still significant.

HOLM PROCEDURE

A more powerful procedure than the Bonferroni adjustment is the one proposed by Holm[3], which is also referred to as weighted Holm-Bonferroni procedure. This procedure is a top-down stepwise calculation whereby the Bonferroni adjustment is recalculated time after time for the comparisons left.[3] For some Authors[5], there would be no need to perform an analysis of variance before the multiple comparisons, provided that the Holm procedure is followed. For the overall cut-off level of significance (usually 0.05) and n multiple comparisons the Holm procedure is as follows:

$$\frac{\textit{Overall cut-off level of P}}{\textit{n - Rank number of significance} + 1}$$

Where:

- Overall cut-off level of P, usually 0.05
- n = number of multiple comparisons
- Rank number of significance obtained sorting the P values from the lowest to the greatest.

Of note, this calculation has to be performed for the P-value of each of the multiple comparisons. In other words, each comparison has its own adjusted P value. The stepwise Holm procedure is also reasonably simple to calculate, but it is more potent than the single-step Bonferroni adjustment, especially when the number of multiple comparisons becomes large[2,4]. The same worked example with 6 multiple comparisons used for the Bonferroni procedure is reported in Table 2 for the Holm adjustments. In this case, after the Holm procedure, 3 comparisons are still significant, showing how this procedure has more statistical power as compared to that of the Bonferroni procedure.

***Table 1.** An example where Bonferroni procedure is applied to a case of 6 multiple comparisons.*

n	Comparsion	P value	Bonferroni procedure for cut-off level, a	Significance
1	A vs. B	0.009	0.050/6 = 0.008	0.009>0.008; NS
2	A vs. C	0.010	0.050/6 = 0.008	0.010>0.008; NS
3	A vs. D	0.001	0.050/6 = 0.008	0.001<0.008; S
4	B vs. C	0.062	0.050/6 = 0.008	0.062>0.008; NS
5	B vs. D	0.120	0.050/6 = 0.008	0.120>0.008; NS
6	C vs. D	0.035	0.050/6 = 0.008	0.035>0.008; NS

Letters **A** to **D** refers to different experimental groups (in a cross-sectional analysis) or time points (in a longitudinal study); **n**, number of comparison; **a**, approximation to 3 decimal places. An overall alpha level of 0.05 has been used in the example. The significance of the comparison after Bonferroni procedure: **NS**, not significant; **S**, significant. Note that all the adjusted **P** values (4th column) for assessing significance are equal.

***Table 2.** The same example in Table 1 where the Holm procedure is applied.*

n	Comparsion	P value (in ascending order)	Holm procedure for cut-off level (a)	Significance
1	A vs. D	0.001	0.05 (6 – 1 + 1) = 0.008	0.001<0.008; S
2	A vs. B	0.009	0.05 (6 – 2 + 1) = 0.010	0.009<0.010; S
3	A vs. C	0.010	0.05 (6 – 3 + 1) = 0.012	0.010<0.012; S
4	C vs. D	0.035	0.05 (6 – 4 + 1) = 0.017	0.035>0.017; NS
5	B vs. C	0.062	0.05 (6 – 5 + 1) = 0.025	0.062>0.025; NS
6	B vs. D	0.120	0.05 (6 – 6 + 1) = 0.050	0.120>0.050; NS

Letters **A** to **D** refers to different experimental groups (in a cross-sectional analysis) or time points (in a longitudinal study); **n**, number of comparison; **a**, approximation to 3 decimal places. An overall alpha level of 0.05 has been used in the example. The significance of the comparison after Holm procedure: **NS**, not significant; **S**, significant. Note that all the adjusted **P** values (4th column) for assessing significance are unequal.

REFERENCES

1. Perinetti G. StaTips Part I: Choosing statistical test when dealing with differences. South Eur J Orthod Dentofac Res. 2016;3:4-5.
2. Glantz S. Primer of Biostatistics. 7th ed. Columbus: McGraw-Hill Education; 2011.
3. Holm S. A simple sequentially rejective multiple test procedure. Scand J Statistics. 1979;6:65-70.
4. Shaffer JP. Multiple hypothesis testing. Ann Rev Psychol. 1995;46:561-84.
5. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs. Holm methods. Am J Public Health. 1996;86:726-8.